A Fast and Flexible Webinterface for Dialect Research in the Low Countries

Roeland van Hout, Nicoline van der Sijs, Erwin Komen, Henk van den Heuvel
CLS, Radboud University
Erasmusplein 1, Nijmegen, the Netherlands
{r.vanhout, n.vandersijs, h.vandenheuvel, e.komen}@let.ru.nl

Abstract

This paper describes the development of webportals with search applications built in order to make the data from the 33 volumes of the Dictionary of the Brabantic dialects (1967-2005) and the 39 volumes of the Dictionary of the Limburgian dialects (1983-2008) accessible and retrievable for both the research community and the general audience. Part of the data was available in a digital format, a larger part only in print. The printed data was semi-automatically converted from paper to structured text (database). This process allowed for streamlining information, applying (semi-)automatic data checks and manually correcting the input. Next, the resulting database was the backbone of a webportal for faceted search requests on the full collection, including filtering and splitting the results on metadata. The design and implementation of the webportals, called e-WBD and e-WLD, are being defined in more detail. The URLs of the portals are: http://e-wbd.nl/ and http://www.e-wld.nl/.

Keywords: web services; data curation; dialects

1. Introduction

The 33 volumes of the Dictionary of the Brabantic dialects (*Woordenboek van de Brabantse Dialecten*, WBD) have appeared in press between 1967 and 2005, while the 39 volumes of the Dictionary of the Limburgian dialects (*Woordenboek van de Limburgse Dialecten*, WLD) were published between 1983 and 2008. The WBD and WLD have been compiled at the Radboud University Nijmegen and at the University of Leuven and both dictionaries started under the guidance of the famous Dutch dialectologist A.A.Weijnen.[1] The Dictionary of the Flemish Dialects (*Woordenboek van de Vlaamse Dialecten*, WVD) was set up according to the same semantic principles, but started later.

The Limburgian dialects are spoken in the provinces of Limburg in the Netherlands and Belgium. The dialects are separated into six dialect areas, as shown in Figure 1.

The same goes for the Brabantic dialects: the dialects are spoken in the Dutch province Northern Brabant, the Belgian provinces Antwerp and Flemish Brabant, and the Brussels-Capital Region, as can be seen on Figure 2. More details can be found in WBD, *part III*, volume *Inleiding & Klankgeografie* (2000).

The WBD and WLD are onomasiologically organized. The main entries are semantic concepts, represented in Standard Dutch (the lemmas). These entries contain the keywords that are dutchified transliterations (normalized spellings) of the dialect forms. The next level gives the phonetic transcription forms (the dialect forms). For each transcription form the location or locations are given where they are used.

Overall, the dictionaries consist of three parts. The first part (WBD 9 volumes, including one devoted to introductory matters; WLD 13 volumes) contains the agricultural terminology. The second part (WBD 9 volumes; WLD 12 volumes), concerns the technical terminology (industries, trades), and finally the third part (WBD 15 volumes, including one devoted to introductory matters; WLD 14 volumes) enumerates the general vocabulary.

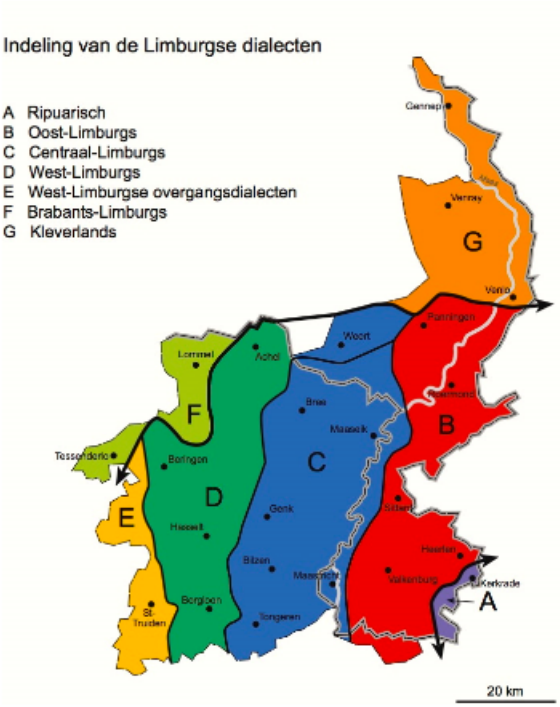

Figure 1. The six Limburgian dialect areas

Unique for Limburg was the mining industry. WLD dedicated a special volume to the vocabulary used in the 19 coal mines in the two Limburgian provinces. The 12 Dutch mines were closed in the 1960s and 1970s, and in the 1990s the 7 Flemish mines followed, making the mine jargon obsolete. In the mines a special, mixed form of dialects was

[1] A complete overview of all volumes can be found at http://www.e-wld.nl/delen and http://e-wbd.nl/delen

spoken: the mine workers came from far and wide and spoke different Limburgian dialects, but also other Dutch, Flemish, German or Walloon dialects, and there were even workers from, among others, Italy, Poland and Brazil. In Dutch mining terminology, German had the greatest influence because many workers came from Aachen and the surrounding area, while French borrowings were frequently occurring in Flemish mines due to the Walloon workers and staff.

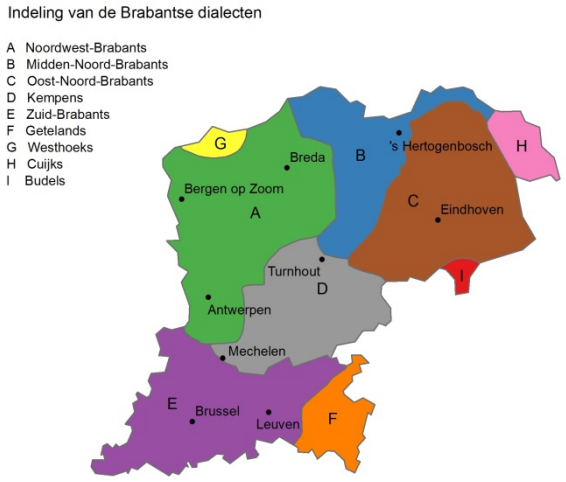

Figure 2. The nine Brabantic dialect areas

The third parts of the dictionaries were prepared by the editors on the computer, using a tailor-made version of FileMaker Pro. These parts have been made digitally available on a website in 2004, as a result of the NWO project D-Square. To this website a cartographic tool was added (Van den Heuvel et al., 2015; de Vriend et al., 2006). No cross-checks were performed on data consistency in D-square and the database only had limited search functions, accessible through a meanwhile fairly obsolete interface.

In 2015 the CLARIN-NL program granted a project called CARE: CurAtion and integration of REgional dictionaries, under supervision of Nicoline van der Sijs and Roeland van Hout. The goal of CARE was to semi-automatically convert text documents of the first two parts of the two dictionaries to structured text (database), and to combine these data with those of the third part. The input format was scanned text, read by optical character recognition (OCR). As output format the Lexical Markup Framework LMF was chosen, as it is an accepted standard in CLARIN for lexical data. In Van den Heuvel, Sanders & Van der Sijs (2016) the methodology used has been described in detail.

The resulting database made it possible to uniform all data and to check information consistency. For instance, there were quite a number of inaccuracies in the so-called Kloeke codes, either in the printed volumes or because of the OCR process. These Kloeke codes, named after the dialectologist G. Kloeke who constructed the system (Kruijsen & Van der Sijs, 2010), refer to locations in the Netherlands and Belgium in a unique way.

The special phonetic symbols turned out to be the hardest problem, particularly for the Limburgian dialects. The editors decided that extremely fine-grained phonetic distinctions between the Limburgian dialects had, for scientific reasons, to be expressed, see Figure 3. Instead of using the IPA for this purpose, they developed an exclusive phonetic script containing many diacritics. The OCR programme could not make head or tail of it, so the resulting text was gibberish. Once the text was converted into a database and all phonetic symbols were put into a single field or column, systematic semi-automatic correction became feasible. This laborious task was done by retired editor Joep Kruijsen.[2] The editors of the Brabantic dialects chose a less elaborate phonetic transcription. This too hampered computer reading, but in a less interfering way.

Along these lines we managed to greatly improve the process of data curation.

1.8.9 VARKENS FOKKEN

(N 76, 37b; monogr.)
[Zich toeleggen op de teelt van varkens.]
fokken: *fǫkə* L 266, 267, 288a, 295, 318d, 328, 371, 376, 0426, Q 16, 27, 32, 112, 117; *fokə* L 320a; **kweken:** *kwēkə* L 381, 423, Q 9; *kwīkə* L 414; *kwekə* P 219, 219a; **baggen kweken:** *bagə kwēkə* L 415, Q 11; **telen:** *tǭlə* L 318d; **vermeerderen:** *vərmiərdərə* L 295; **varkens optrekken:** *vɛrkəs optrɛkə* L 265; **trekken:** *trękə* Q 197; **tuchten:** *tsøxtə* Q 117; *tsyxtə* Q 121; **varken houden:** *vɛrkən hǭgən* Q 1; *vɛrkə hā* P 188.

Figure 3. An example of the phonetic symbols used in WLD

All resulting LMF-files were stored in open access at the Meertens Institute in Amsterdam (which is a CLARIN Data Centre).

The next step was to design and implement a webportal. The idea was that a complete, curated database with a transparent webinterface providing precise and powerful selection and search tools would open up new avenues for dialect research in the Brabant and Limburg areas, in particular because the data are there for everyone on the internet. In the following sections we describe how we dealt with the design and implementation of the

[2] These corrected data have not yet been made available through the webportal, but this will be done in the future.

webportals, now accessible at http://e-wbd.nl/ and http://www.e-wld.nl.

2. Set-up of the Webportal

Figure 4 shows the main elements of a dictionary entry as presented in the volumes.

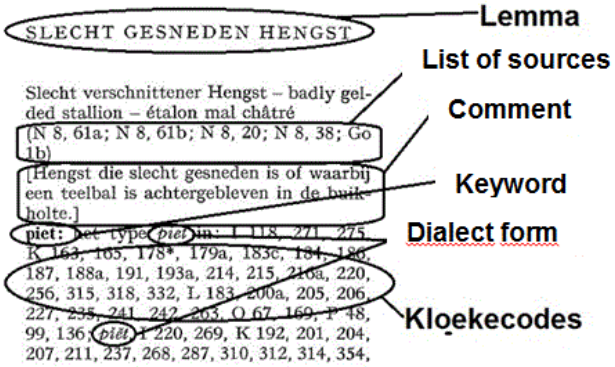

Figure 4. Example from the dialect dictionary with explanation of how it is structured.

The main components contained in the database are therefore:

- Lemma (= concept)
- Lemma comment
- Source list
- Keyword (= dialect entry in normalized spelling)
- Dialect form (= phonetic spelling of dialect form)
- Kloeke code and associated place name

At the start of the webportal project we set up a list of requirements which the portals should fulfill. The webportals should contain:

- Information about the digitization process of the paper books into a database
- A map of the Limburgian and Brabantic areas or dialects (see Fig. 1 and 2)
- Full overview of the volumes
- Access to the PDF versions of all volumes
- Overviews of all lemmas, keywords, places
- Search options at the level of lemma, keyword, place, allowing
    - Wildcards in the search terms
    - Filter options for place, Kloeke code, volume
- Hyperlinks between retrieved lemmas and corresponding keywords
- Access to the PDF files of the books via the results of the query.

3. Implementation of the Webportal

The e-WLD and e-WBD webportals are written in Python 3.5, building on the freely available Django web application framework.[3] Django is a Python-based framework that makes it easier to build web applications quickly and with less code. The code of the portal has been developed using the PTVS facilities of Microsoft Visual Studio 2013.[4] The site is served as an uwsgi application that runs under Apache on one of the webservers at Radboud University Nijmegen. All entries (close to 2 million dialect words) are stored in a Sqlite database.[5]

The map on the site's homepage, is based on the most recently available geographical data, and it shows the Belgian and Dutch dialect areas that are covered by the WLD.

The webportal allows listing the available dialect forms based on their lemma (*begrippen* in Figure 5), or on their keyword (*trefwoorden* in Figure 4). It is also possible to obtain a list of the locations (*plaatsen* in Figure 5) or (for WLD) a list of the coal mines where dialect forms have been collected.

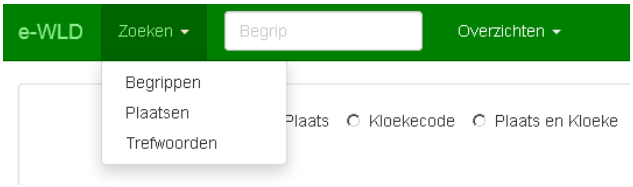

Figure 5. The menu bar's basic search functions

The web application's menu bar (Figure 5) provides access to three basic search functions: search a lemma or concept (e.g. "aard*"), search a keyword (e.g. "*appel") or look for a location (e.g. "kern*"). The result of a lemma search is a list of all the keywords, dialect forms and locations where the lemma is found; searching for a keyword lists all the lemmas, dialect forms and places where the keyword is found. A search for a location yields all the different locations that match the query, and a list of lemmas, keywords and dialect forms for each of the matching locations. Lemma's and keywords that are part of a search's result list are clickable: they allow fast access to the respective lemma and keyword entries.

Two of the search functions allow making use of additional filtering. The search for a lemma can be extended by filtering for: Kloeke code (a location code for the Netherlands and Belgium; see above), name of the location, dialect form, publication volume and, when applicable and only for WLD, the name of the mine where words have been elicited. Searching for a keyword adds a filter for the lemma and one for keyword comments. Search results can be downloaded as tab-separated text, as an Excel file or as an HTML file.

Figure 6 illustrates one type of search result. Looking for the keyword 'inktappel' yields two lemmas that contain this keyword: *dennentakje met een harsknopje* 'little pine branch that has a resin bud' and *galnoot* 'gall apple'. Each of these lemma results is accompanied by a list of dialect forms, and each dialect form contains a list of locations (and possibly mines) where this form has been

[3] See http://www.djangoproject.com.

[4] PTVS is the Python toolkit for Visual Studio. The sources of the webportal are available at https://github.com/ErwinKomen/RU-wld.

[5] The development process required a number of attempts to import all the data, which is why a special asynchronously running admin component was added, allowing the import process to be monitored. Importing the data from just one of the published volumes requires one to two hours.

attested. The locations can be accompanied by location-specific comments. The list of results is finished by a list of volumes where these lemmas occur.[6]

Figure 6. Results obtained for the keyword 'inktappel'

An interested researcher can, on the basis of the search results, access the PDF version of the original publication and check the text over there.

4. Data model

The data model that is being used to represent the information in the dialect dictionary is shown in Figure 7.

The basic unit in the data model is the `Entry`. Each entry contains the information pertaining to a dialect word: the word itself ('woord' in the model), additional information to this particular word ('toelichting') and a number of links to other parts of the database:

- A link to the `Lemma` of which this word is part
- A link to one of the `Dialect`'s where words have been elicited
- A link to the keyword (`Trefwoord`) it belongs to
- An optional link to a number of mines (`Mijn`) this word has been found (only for WLD)
- A link to the published edition (`Aflevering`) where this particular combination of dialect-word/keyword/lemma can be found.
- Each edition links to one of the three parts (`Deel`) to which the editions belong.

The data model chosen has direct consequences for processing and using the database, since the distance between which any two elements in a relational database can be found determines the complexity of a search and, consequently, the speed of searching. All `Entry` elements of one `Lemma` are found fast enough, but retrieving and sorting the related keywords (`Trefwoord`) is more complex. The use of Django, however, takes care of optimalizations behind the scene, arriving at a workable database.

5. Conclusions and Future work

The e-WLD website was launched on December 17, 2017 (Van der Sijs 2017) and was positively reviewed in local and national Dutch news papers. The same holds for the e-WBD website, which was launched on 14 December 2017.

For the first time we can count how much material the dictionaries actually contain: in the e-WBD there appear to be 15,794 concepts, 140,091 keywords and 1,704,116 dialect forms, collected in more than 1000 dialects (each place/location representing its own dialect). In the e-WLD there appear to be 17,539 concepts, 137,231 keywords and 1,759,090 dialect forms, collected in more than 1000 dialects (each place/location representing its own dialect).

Answers on many questions can now for the first time be given, for instance whether a specific word for a certain concept is really unique or not and where particular word forms can be found in the area. Another relevant research question is the relationship between word form distributions and semantic concepts. Word lists per location can be made now quite easily, which is highly supportive for writers of local dialect dictionaries. People can easily check whether a specific word form with a specific meaning was previously documented for a Limburgian or Brabantic dialect, or perhaps in another Limburgian or Brabantic dialect. Both webportals seem to meet both professional and popular needs of people interested in dialects.

Several collaborations have been set up in order to make sure that the same uniform data base model is used among various projects dealing with these dictionaries, including the Dictionary of the Flemish dialects (*Woordenboek van de Vlaamse Dialecten*, WVD).

The University of Ghent now hosts the overarching project 'Dictionary of the Southern Dutch dialects. An integrated lexicological infrastructure for the Southern Dutch dialects' (DSDD). Aim of this project is to integrate and standardize all three southern comprehensive dialect lexicographic databases (Limburgian, Brabantic, Flemish). The consortium involved includes linguists, ICT support staff, digital humanities experts and geographers. This project will be carried out in close co-operation with the INT, the Institute for the Dutch Language in Leyden.

Finally, the overarching data model is designed in such a way that not only other regional onomasiological dictionaries can be added, but also local semasiological dictionaries, i.e. dictionaries that contain a description of the dialect vocabulary of a specific place or small region. For the Netherlands, these dictionaries are collected, digitized and curated at the Meertens Institute, and made available through http://www.meertens.knaw.nl/ewnd/. The Flanders counterpart can be found at https://www.woordenbank.be.

[6] The abbreviated publication denotation consists of the part number (running from I to III) followed by the volume number. Part III deviates from this scheme, having each publication denotation consisting of: part number (that is: III), division number (ranging from 1-4) and then the volume number.

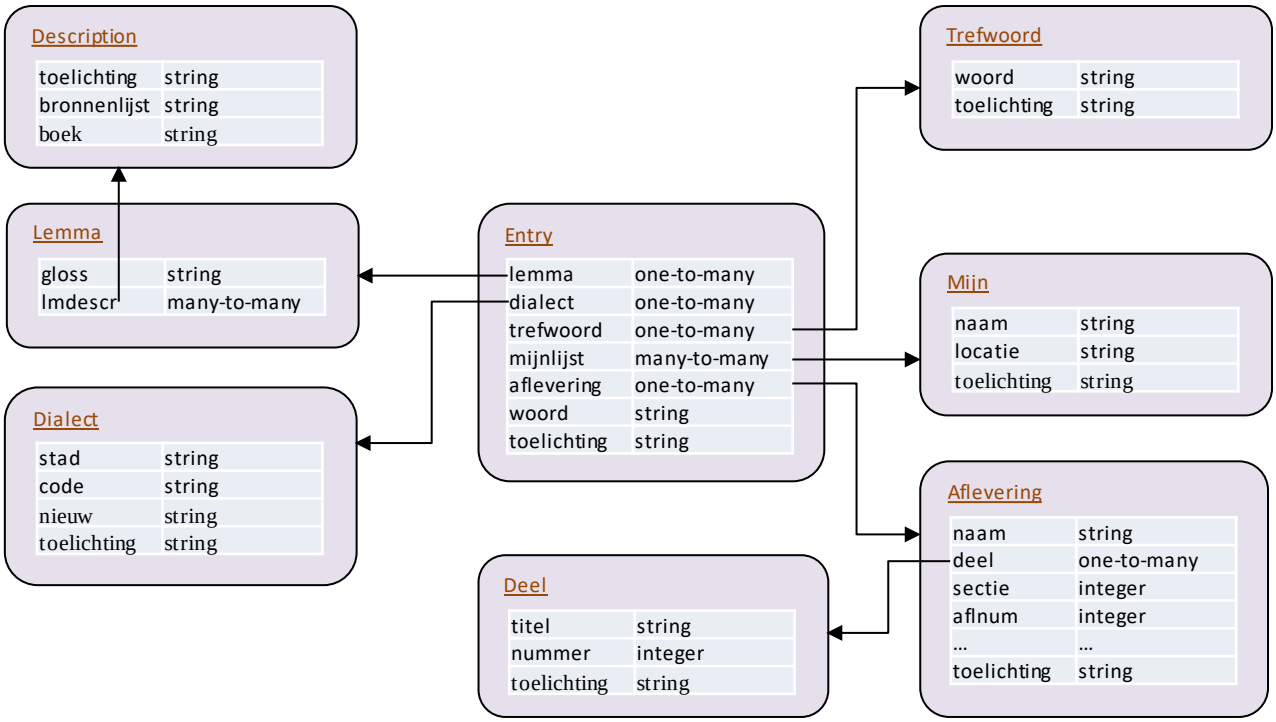

Figure 7. The data model used for the webportals of the dialect dictionaries

6. Acknowledgements

The data form the core of the web application. The data curation has been executed by former editor Joep Kruijsen, research assistants Linda van Meel and Aukje Borkent, student assistants Jorik van Engeland, Inge Otto, Lisette van der Heijde and Hanna van den Heuvel, interns Maaike Borst and Eline Dimmendaal. We greatly thank them for their meticulous and diligent work on the manual text processing. We also thank volunteers Jantien Kettenes-Van den Bosch and Herman Wiltink for their work on the preprocessing of the text documents. We thank the *Raod voor 't Limburgs* and the chair of *Diversity in Language and Culture* of Tilburg University for their moral and financial support, and Thijs Hermsen for drawing the website's logos showing the Limburgian and Brabantic dialect areas.

7. Bibliographical References

Heuvel, H. van den, Oostdijk, N., Sanders, E., and Lint, V. de (2015). Data curations by the Dutch data curation service. Overview and future perspective. In: Odijk, J. (2015). *Selected Papers from the CLARIN 2014 Conference*, October 24-25, 2014, Soesterberg, The Netherlands., pages 54–62. Linkoping Electronic Conference Proceedings, 116. http://www.ep.liu.se/ecp_article/index.en.aspx?issue=116;article=005.

Heuvel, H. van den, Sanders, E. and Sijs. N. van der (2016). Curation of Dutch Regional Dictionaries. *Proceedings LREC 2016* (10th edition of the Language Resources and Evaluation Conference, 23-28 May 2016, Portorož (Slovenia).

Kruijsen, J. and Sijs, N. van der (2010). Mapping Dutch and Flemish. In: Lameli, Alfred, Kehrein, Roland and Rabanus, Stefan (eds). *Language and Space: An International Handbook of Linguistic Variation: Language Mapping, Handbooks of linguistics and communication science*; 30.2, pages 180–202. De Gruyter Mouton.

Sijs, Nicoline van der (2017). Limburg heeft het! Het Woordenboek van de Limburgse Dialecten digitaal doorzoekbaar. In: *Neerlandia* 1, pages 41-43.

Sijs, Nicoline van der (2018). Onverwoestbaar Brabants. Over het elektronische Woordenboek van de Brabantse Dialecten. In: *Neerlandia* 1.

Vriend, F. de, Boves, L., Heuvel, H. van den, Hout, R. van, Kruijsen, J., and Swanenberg, J. (2006). A unified structure for Dutch dialect dictionary data. In : *Proceedings LREC 2006*, Language Resources and Evaluation Conference, Genova, Italy 2006, pages 1660–1665.

WBD = *Woordenboek van de Brabantse Dialecten.* Assen/Maastricht/Groningen/Utrecht, 1967-2005.

WLD = *Woordenboek van de Limburgse Dialecten* (WLD). Assen/Maastricht/Groningen, 1982-2008

WVD = *Woordenboek van de Vlaamse Dialecten*, Tongeren, 1979-.

Wijngaard, T. van de, and Keulen, R. (2007). De indeling van de Limburgse dialecten. In: Keulen, R. Wijngaard, T. van de, Crompvoets, H. and Walraven, F. (ed.). Riek van klank. *Inleiding in de Limburgse dialecten*, Sittard, pages 15-22.